Lab 4: Model selection for the FEV data

Create a short reproducible report answering the questions below, The report should be less than 3 pages, including all figures, and should be submitted as both PDF and Rmd formats. You do not need to show your code in the PDF report.

Data exploration

This lab will cover model selection and checking for multiple linear regression using the FEV dataset (found in the Vanderbilt Datasets repository). This dataset contains 654 observations on 6 variables, including a continuous outcome variable, forced expiratory volume (FEV) which is a measure of lung health and strength. To begin, load (install if necessary) the Hmisc package to load the dataset from the Vanderbilt website. The brief FEV dataset codebook can be found on this webpage.

```
library("Hmisc")
getHdata(FEV)
```

Exercise 1 Before looking at the data, what types of relationships do you expect to see between each of these variables and FEV? Justify your answers briefly.

Exercise 2 Think about the different measures variables that you have at your disposal. Hypothesize at least two possible interactions, and come up with a plausible justification about why such an ineraction might exist.

Exercise 3 Generate a few simple plots of the data to evaluate the possible bivariate relationships that exist. Based on these plots, do you think that linear relationships will be sufficient to explain the data?

Recall (from lecture 5) that when you have a large sample size relative to the number of possible covariates (which we do), one justifiable way to build a model is to

- Include key covariates of interest.
- Include covariates needed because they might be confounders.
- Include covariates that your colleagues/reviewers/collaborators will demand be included for face validity.
- Test a reasonable/limited number of interaction terms if it seems appropriate.

Exercise 4 Before fitting any models, identify somewhere between 2 and 6 multiple linear regression models that you think follow the rules above, i.e. write down what each model is, and which covariates and/or interactions it includes. Fit each model. Compare the coefficients from each of the models as well as the adjusted R^2 values. For the 2-3 models with the best adjusted R^2 values, look at residual plots. Do you see any worrisome patterns that suggest the model assumptions aren't met? If so, do something to try to improve it.

Exercise 5 Based on your work above, pick one model that represents what you think is the "best" model for this data. Justify your choice, and interpret what you think are the most important or interesting relationships with the covariates and FEV. Illustrate one of these relationships with a figure. If you were to collect similar data like this in a new study, what additional variables would you be interested in collecting and why?